Analysis of Shifts & trends of Organizations in Indonesia using Tweets & RSS feeds

by

Sathishkumar Poornachandran

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2013 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Arunabha Sen
Mark Woodward

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

With the advent of social media (like Twitter, Facebook etc.,) people are easily sharing their opinions, sentiments and enforcing their ideologies on others like never before. Even people who are otherwise socially inactive would like to share their thoughts on current affairs by tweeting and sharing news feeds with their friends and acquaintances.

In this thesis study, we chose Twitter as our main data platform to analyze shifts and movements of 27 political organizations in Indonesia. So far, we have collected over 30 million tweets and 150,000 news articles from RSS feeds of the corresponding organizations for our analysis. For Twitter data extraction, we developed a multi-threaded application which seamlessly extracts, cleans and stores millions of tweets matching our keywords from Twitter Streaming API. For keyword extraction, we used topics and perspectives which were extracted using n-grams techniques and later approved by our social scientists. After the data is extracted, we aggregate the tweet contents that belong to every user on a weekly basis. Finally, we applied linear and logistic regression using SLEP, an open source sparse learning package to compute weekly score for users and mapping them to one of the 27 organizations on a radical or counter radical scale. Since, we are mapping users to organizations on a weekly basis, we are able to track user's behavior and important new events that triggered shifts among users between organizations. This thesis study can further be extended to identify topics and organization specific influential users and new users from various social media platforms like Facebook, YouTube etc. can easily be mapped to existing organizations on a radical or counter-radical scale.

Dedicated to my mom and dad

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my advisor Dr. Hasan Davulcu for providing an excellent infrastructure to work on my thesis. I have been privileged and fortunate to work under him for more than a year. This thesis work would not have been possible without his guidance, encouragement and motivation.

I would like to thank Dr. Arunabha Sen and Dr. Mark Woodward for accepting my request to be a part of my thesis committee and also for the guidance.

I would also like to thank Nyunsu Kim for providing slides and guiding me during my work in the Cognitive Information Processing System (CIPS) lab. I would also like to thank all the members of CIPS Research lab for providing valuable insights in some way or the other.

Personally, I would like to thank my family members for their guidance and support.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Social media websites (like twitter, Facebook etc.,) have created a public space on online debates and social issues [1]. When an important incident happens in some part of the world, we could see people sharing their opinions, sentiments and sometimes taking perspectives in a hot debate. Information thus gathered from a debate can be crucial to track the behavior of an individual or a political group.

In this thesis, we developed an end to end framework to analyze shifts and behaviors of various users and organizations using tweets and documents extracted from twitter streaming API and RSS feeds of the respective organizations. Initially, we crawled over 37,770 documents (news articles, events etc.,) from 27 different organizations in Indonesia and built our training model using linear (for Individuals) and logistic regression (for groups). Once our training model is built, we started extracting real time tweets from Twitter streaming API with the help of top K matching keywords that were previously extracted using various techniques explained in [1]. Finally, we aggregated all the users tweet on a week basis and computed weekly scores. With the help of the generated scores, we mapped every user to an organization on a weekly scale. Since we track individuals on a weekly scale, we are able to study their patterns and radical behaviors over a period of time and track important news and events on the way.

Rest of the thesis work is arranged as follows: Chapter 2 discusses about the various components and the overall architecture of the system in detail. Chapter 3 illustrates the experiments and results. Chapter 4 has the scenarios. Chapter 5 discusses the summary and chapter 6 covers the future study and improvements of the system.

Chapter 2

SYSTEM ARCHITECTURE AND COMPONENTS

This chapter deals with the proposed architecture and various components of our system. It can be divided into

1) System Architecture.
2) Data Collection.
3) Filter Analysis.
4) Data Cleaning.
5) Data Aggregation.
6) Data Classification.

2.1: SYSTEM ARCHITECTURE:

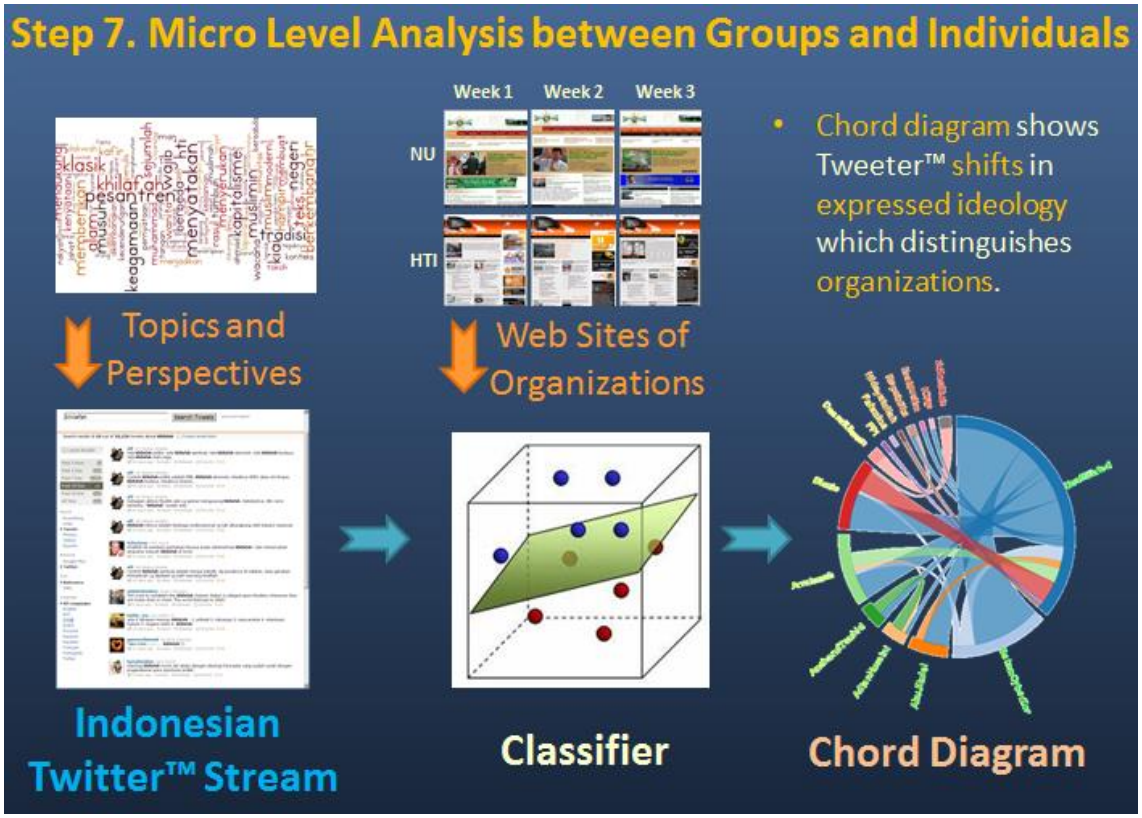

Figure 2.1: Overall system architecture (Image copied from [10])

The overall architecture of our system can be studied from Figure 2.1. Initially, we collected documents (new articles, events etc.) for building our training dataset from the list of organizational websites provided by our experts. In addition to that, we subscribed to Twitter Streaming API and started extracting real time tweets with the help of topics and perspectives as keywords. The extracted tweets were then cleaned, tokenized, aggregated and sent to the classifier for computing weekly and organization score for every user. The classifier was previously built using linear and logistic regression model. The generated scores were then sent to the chord diagram for data visualization.

2.2. DATA COLLECTION

We collected data from the official websites of the political organizations, Twitter Streaming API and subscribed to RSS news feeds.

2.2.1. ORGANIZATIONAL WEBSITES

Initially, we found 27 different organizations in Indonesia with the help of our social scientists from Indonesia and labeled them as radical or counter radical organizations. For the sake of simplicity, we made a naive assumption that documents crawled from a radical organization would also be radical. On a similar note, documents crawled from a counter-radical organization would also be counter-radical. Since, every website has their own markups, we wrote site-specific crawlers and downloaded 37,770 different documents in the form of news articles, events, publications etc. Keywords from these documents had formed the base for our training model which was explained in section 2.4. Table 1 shows the list of organizations, their radical index and the total number of documents crawled from the respective organizations.

ORGANIZATION	R/CR	Count
AbuJibriel	R	9
AdianHusaini	R	75
AnsharutTauhid	R	47
Arrahmah	R	2708
DaarulUluum	CR	61
EraMuslim	R	5413
Fahmina	CR	722
FPI	R	197
HizbutTahrir	R	1871
ICRP	CR	126
Interfidei	CR	31
IslamLiberal	CR	893
Lakpesdam	CR	243
LKIS	CR	58
MaarifInstitute	CR	279
MillahIbrahim	R	77
MMJabodetabek	R	37
Muhammadiyah	CR	298
NU	CR	23137
Paramadina	CR	17
PKS	R	51
PPIM	CR	57
WahidInstitute	CR	502
Hidayatullah	R	561
ICDW	CR	100

IkhwanWeb	R	100
DewanDakwah	R	100

Table 2.1: List of organizations, R/CR and number of documents crawled

2.2.2. TWITTER STREAMING API:

For our analysis, we collected 16 weeks of tweets from Indonesia between October 10, 2012 and January 29 2013 by applying keyword and location filters. In this timespan, we received 15,320,173 tweets with 2,880,293 unique users. To handle data of such scale, we used Thread pooling to extract multiple tweets at the same time. The extracted tweets were parsed and stored into a relational database for persistence. The schema design is given below.

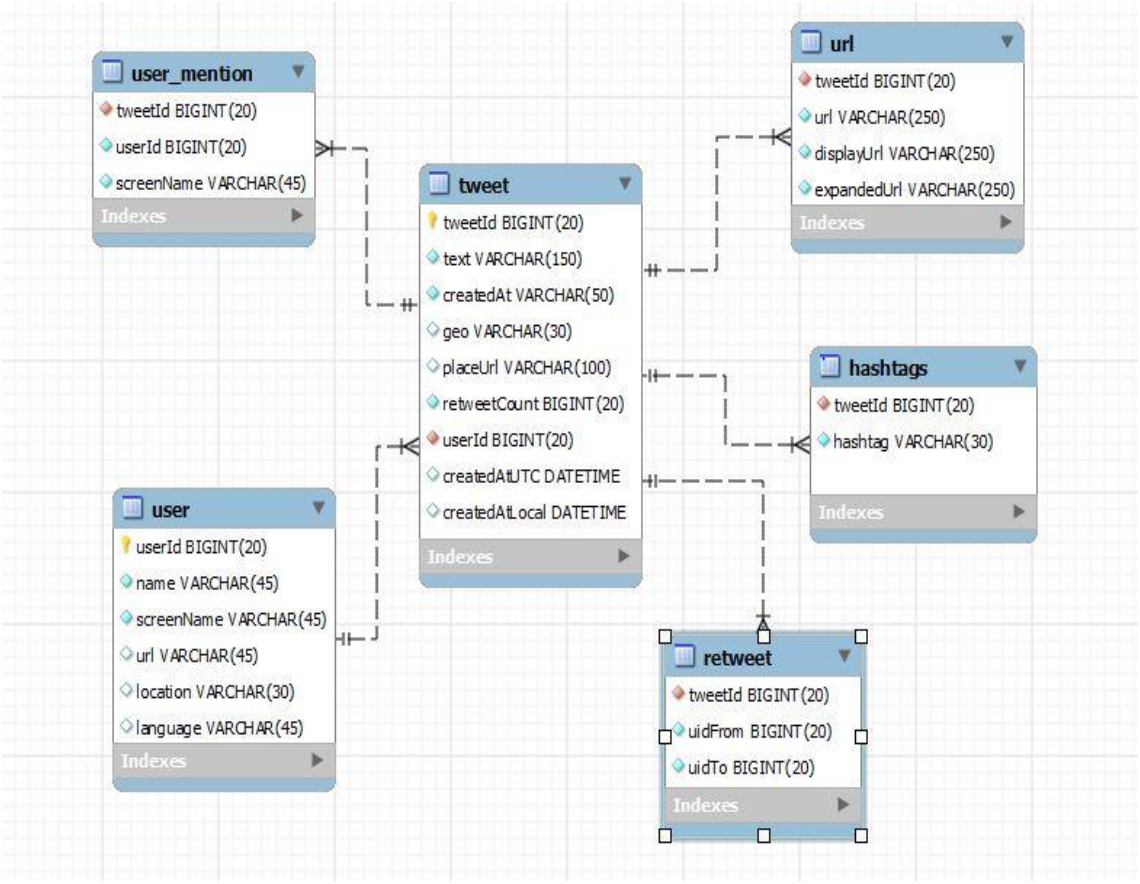

Figure 2.2: Database schema design for storing tweet contents

2.2.3 RSS FEEDS

For regularly updating our training dataset, we subscribed to official RSS feeds from the organizations. So far, we have accumulated over 200,000 articles through 16 RSS feed urls.

2.3 FILTER ANALYSIS

Keywords were used for filtering tweets from Twitter Streaming API. We generated candidate list of topics and perspectives using term-frequency and inverse document frequency techniques [1] [2]. We later asked our social scientists to identify the most important keywords from the above candidate list. Finally, we came up with a list of 29 & 26 radical and counter radical keywords respectively [1]. The above 55 (29+26) keywords in addition to 27 organization names formed the base of our keyword filtering. The candidate lists of keywords (separated by comma) are shown below.

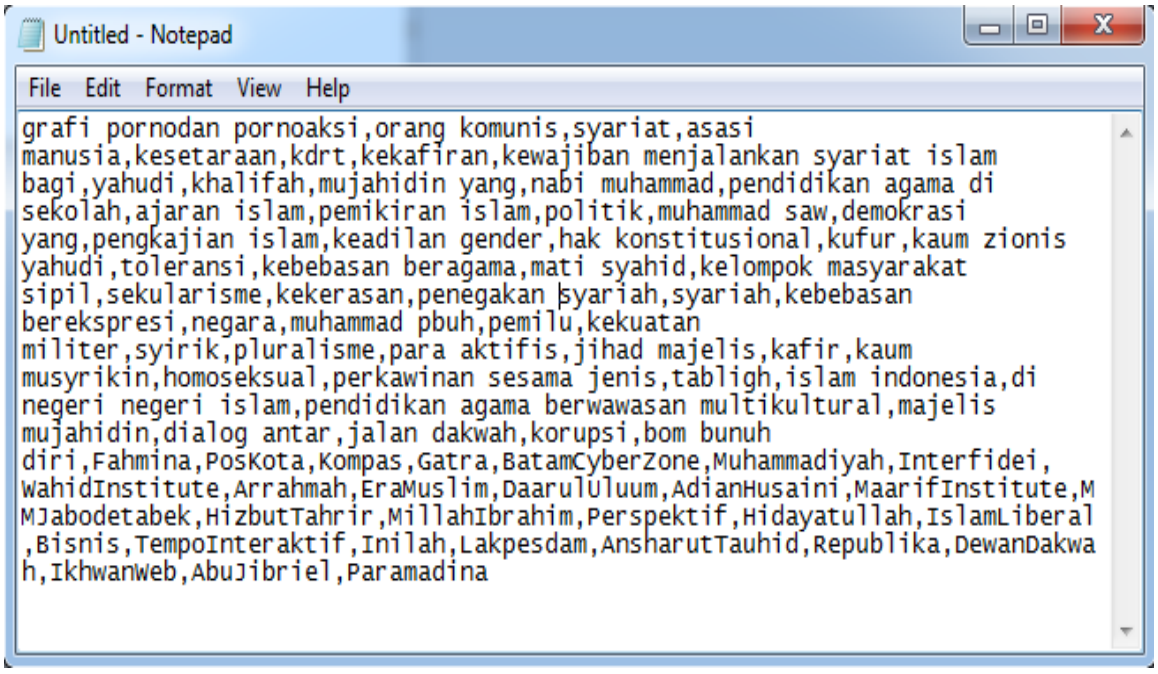

Figure 2.3: List of candidate keywords used for filtering tweets

2.4 DATA CLEANING

Data cleaning is one of the most important modules in our data processing stage as irrelevant data can grossly bring down the accuracy of our classifier. Data cleaning mainly constitutes document, URL and tweet cleaning.

2.4.1 DOCUMENT CLEANING

Since our documents were used as a training data set, it was imperative to clean it to make our predictions more accurate. The extracted articles were in the form of HTML pages with multiple markup tags intertwined. We had to skip most of the information in the HTML pages and extract the original articles with publication date and author information if possible. For extracting the article text from HTML pages, we used an open source Java library “Boilerpipe” [3]. Boilerpipe uses shallow text features to extract article contents and it was discussed in [4].

2.4.2 TWEET CLEANING

As far as tweets were concerned, we had to deal with huge number of “Twitter bots” which decreased the accuracy rate. “Twitter bots” are the spam accounts that try to get you to click on spam links [5]. Some of the efficient ways to reduce the threats were briefly described in [5].

2.4.3 URL CLEANING

Extracted tweets contained millions of useful urls which were used alongside tweet contents. We mainly focused on urls that contained news articles, important events, perspectives etc. and removed most of the spam and home urls. A tweet containing more than one url has a very high probability of having a home page url. We programmatically removed all these home urls by counting the number of forward slashes in a given url. If the total number of forward slashes is less than 4, we assumed that it is a home page url.

2.5 DATA AGGREGATION

In the previous section, we discussed about extracting, filtering and cleaning the user contents. After preprocessing the data, we stored all the contents (tweets, user information, urls etc.) in a normalized database. If we closely look at the database schema, there exist a 1-1 correspondence between users and tweet contents. Since a single tweet (just 140 characters) does not provide much information about a user's perspective, we aggregated all possible user contents on a weekly basis. Hence, we merged all the tweets, title and body contents from the tweet urls with respect to a twitter user on a weekly basis. By this approach, we aggregated enough information about all the twitter users over a period of time. Since we have too many users on any given week, we created a multi-threaded client application to handle data of such scale. In the multi-threaded environment, each thread establishes a separate connection with the database, fetches the entire user contents seamlessly and stores it back to the database on a separate table. A simple demonstration is shown below.

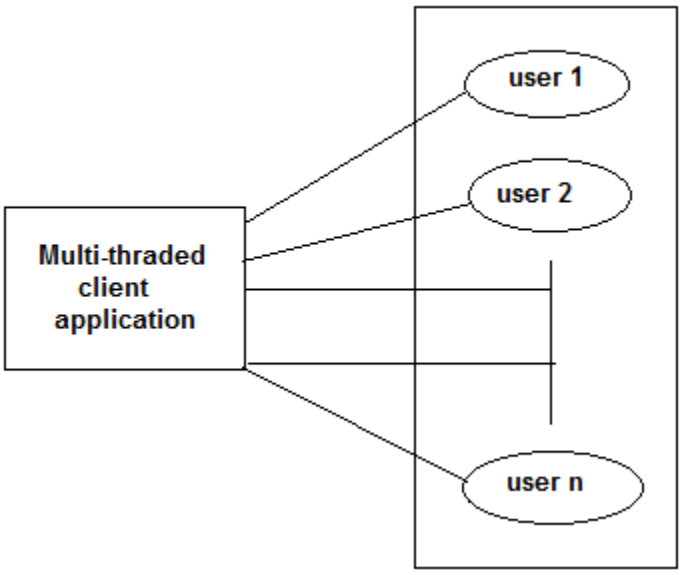

Figure 2.4: Multi-threaded Architecture to extract weekly user contents

2.5.1 TWEET EXTRACTION

Tweet extraction was done by simply collecting each user keywords (from tweet) on a given week.

2.5.2 TITLE EXTRACTION

Title extraction was done by extracting the urls from user tweets. Once the url was parsed from tweets, we extracted the HTML contents from it. The HTML content was then sent to title extractor. The title extractor uses “Longest common subsequence problem” to compare the body and the title of a given HTML page to extract the title. “Longest common substring problem uses dynamic programming technique to find the longest string that is a substring of given string” [6]. We used the same algorithm given in [6].

```
function LCSubstr(S[1..m], T[1..n])
    L := array(1..m, 1..n)
    z := 0
    ret := {}
    for i := 1..m
        for j := 1..n
            if S[i] = T[j]
                if i = 1 or j = 1
                    L[i,j] := 1
                else
                    L[i,j] := L[i-1,j-1] + 1
                if L[i,j] > z
                    z := L[i,j]
                    ret := {S[i-z+1..i]}
                if L[i,j] = z
                    ret := ret ∪ {S[i-z+1..i]}
            else L[i,j]=0;
    return ret
```

Figure 2.5: Algorithm for longest common substring (Note: Image copied from Wikipedia page [6])

2.5.3 ARTICLE EXTRACTION

As we have discussed in section 2.3.1, any HTML document can be sent to Boilerpipe [3] and extract the actual text contents from it. To increase the accuracy of the classifier, we considered only the first 50 words from the article. Here, we made an assumption that gist of the article can be found in the first 50 words itself. We have also considered using the commercial version of alchemy API [7] for extracting important keywords from HTML documents. We persisted with Boilerpipe as we have dealt with millions of HTML documents on a daily basis.

2.5.4 STOP WORD ELIMINATION

After we had extracted all the contents (discussed in section 2.4.1, 2.4.2, and 2.4.3) of every user on a weekly basis, we merged the contents and sent it to a stop word eliminator. The stop word eliminator uses a unique file list to match keywords and eventually eliminates them. The file list contains 485 Indonesian stop words collected over a period of time with the help of experts and social scientists.

2.5.5 FINAL THOUGHTS ON DATA AGGREGATION

In this chapter, we discussed about extracting and merging weekly contents from all the twitter users. The extracted tokens can directly be used to compute weekly scores for the users and eventually classify them on a radical or counter radical scale. In addition to aggregating contents from different sources, we have to clean it beforehand and ensure that proper contents are getting tokenized. For tokenization, we discussed about using stop word eliminators to keep aside the non-contributing keywords as it decreases the efficiency of the classifier. The numerical information of the users and their contents were discussed in section 3.

2.6 DATA CLASSIFICATION

In section 2.5, we discussed about integrating all the contents that belong to weekly twitter users. In this section, we will discuss about classifying the users on a radical or counter-radical scale using linear and logistic regression. For this, we used an open source package SLEP [11], an open source sparse learning package to compute weekly score for users and organization score for every organization. Based on the generated score, we mapped the users to one of the organizations in Indonesia.

2.6.1 TRAINING DATA COLLECTION

As we discussed in section 2.2.1, we initially found 27 radical and counter radical organizations in Indonesia with the help of our experts and social scientists. We then created site specific crawlers for each organization, extracted news, articles, events etc., and labeled those documents as radical or counter radical based on a naive assumption that documents crawled from a radical organization and counter radical organization must also be radical and counter radical respectively. Details of organizations, exact number of documents crawled were given in detail in section 2.1.1.

2.6.2 LINEAR REGRESSION MODEL

We defined our training model using the training datasets in a general sparse learning framework since the vocabulary of the corpus is much larger than an individual document aggregation of keywords [12]. We tried to solve L1 regularized least squares problem given below.

$$\min_x \frac{1}{2} \|Ax - y\|_2^2 + \frac{\rho}{2} \|x\|_2^2 + \lambda \|x\|_1$$

Figure 2.6: LeastR optimization problem (Note: Image copied from [11])

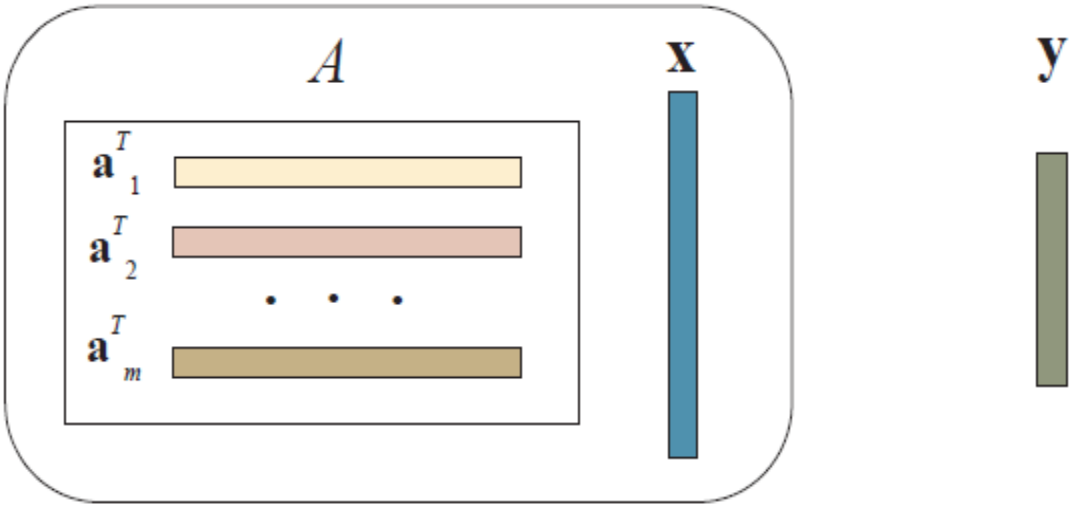

Figure 2.7: Document -> Term matrix arrangement (Note: Image copied from [11])

Where A is a Document -> Term sparse matrix where document is individual document and Term is the overall vocabulary in the document corpus excluding the non-contributing stop words [11] [12].

Y is the class variable {+1, -1}. We mark the radical document as +1 and counter radical organization as -1. This is based on the radical and counter radical information provided by our experts on the organizations [12].

X is the resultant variable that gives the keyword scores for all the terms in the corpus. Based on this X vector, we compute the weekly score for the users. We used different Lambda value given in Figure 2.6 to optimize X vector.

2.6.3: WEEKLY SCORE COMPUTATION

Using the training model generated in section 2.6.2, we calculated the scores for every user on a weekly basis. This is possible by simply multiplying the resultant vector X with the real time document-term matrix given in Figure 2.8. In this scenario, the document-term was generated from tweet users' aggregated weekly contents. If the generated score is greater than 0, we

classify the users as radical users. On the contrary, if the score is less than 0, we classify the users as counter radical users.

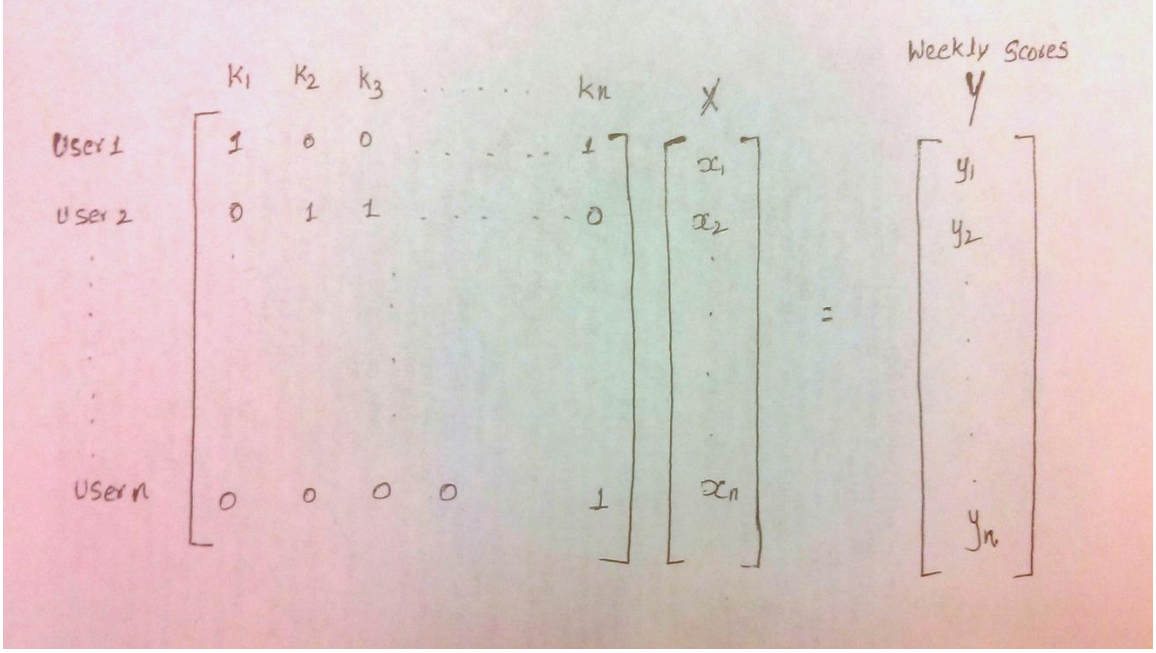

Figure 2.8: Weekly score computation

Chapter 3

EXPERIMENTAL RESULTS AND ANALYSIS

This chapter illustrates the experimental results, tweet statistics and analysis. As we have discussed in the section 2.1.2, we collected 15,320,173 tweets with 2,880,293 users over a period of 16 weeks (October 10, 2012 and January 29 2013).

3.1 WEEKLY TWEET STATISTICS

The below table and the graph chart have the tweet statistics of users from Indonesia on a weekly basis. The last column shows the number of users with 7 or more tweets.

Week	Start date	End date	Total tweets	Total USER	Users with 7 or more tweets
1	10/10/2012	10/16/2012	635537	297867	3089
2	10/17/2012	10/23/2012	341672	174811	3317
3	10/24/2012	10/30/2012	591029	307414	3275
4	10/31/2012	11/6/2012	389305	195923	3675
5	11/7/2012	11/13/2012	829612	387498	7291
6	11/14/2012	11/20/2012	1424298	556131	15534
7	11/21/2012	11/27/2012	1293938	517951	14930
8	11/28/2012	12/4/2012	830239	335636	7832
9	12/5/2012	12/11/2012	1259696	537304	9900
10	12/12/2012	12/18/2012	1263102	536494	9950
11	12/19/2012	12/25/2012	710778	368371	6596
12	12/26/2012	1/1/2013	1203601	539485	7837
13	1/2/2013	1/8/2013	836599	384639	7717
14	1/9/2013	1/15/2013	1087851	488371	9684
15	1/16/2013	1/22/2013	954230	453334	8656
16	1/23/2013	1/29/2013	1668686	739772	

Figure 3.1: Tweet statistics on a weekly basis

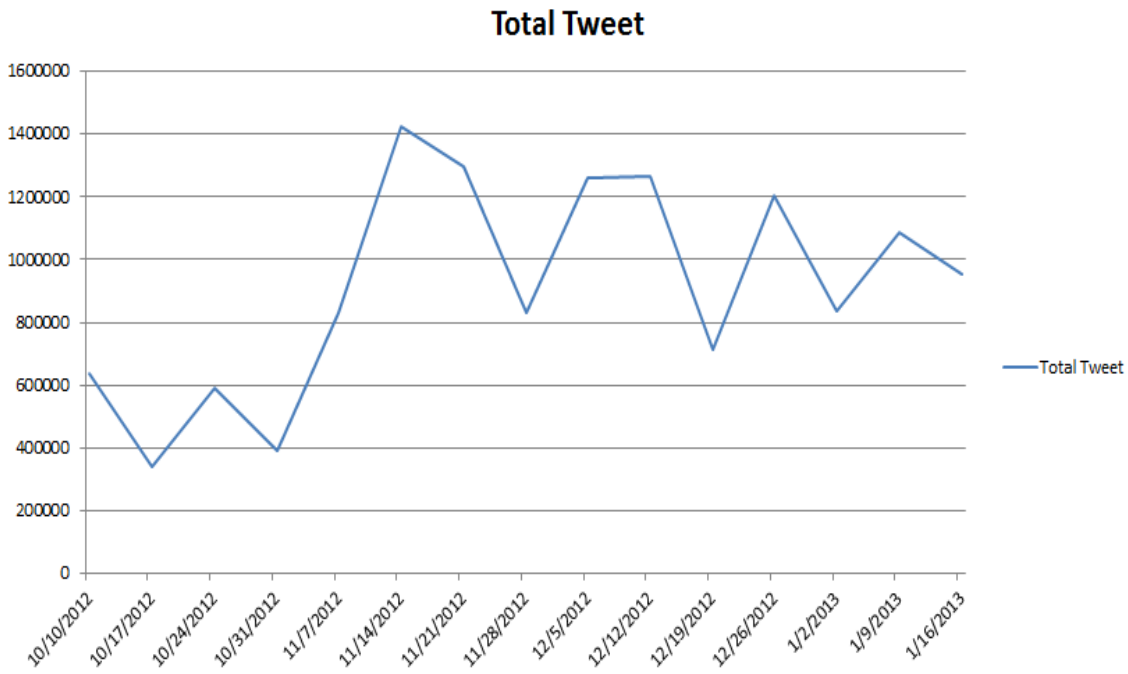

Figure 3.2: Total number of tweets extracted

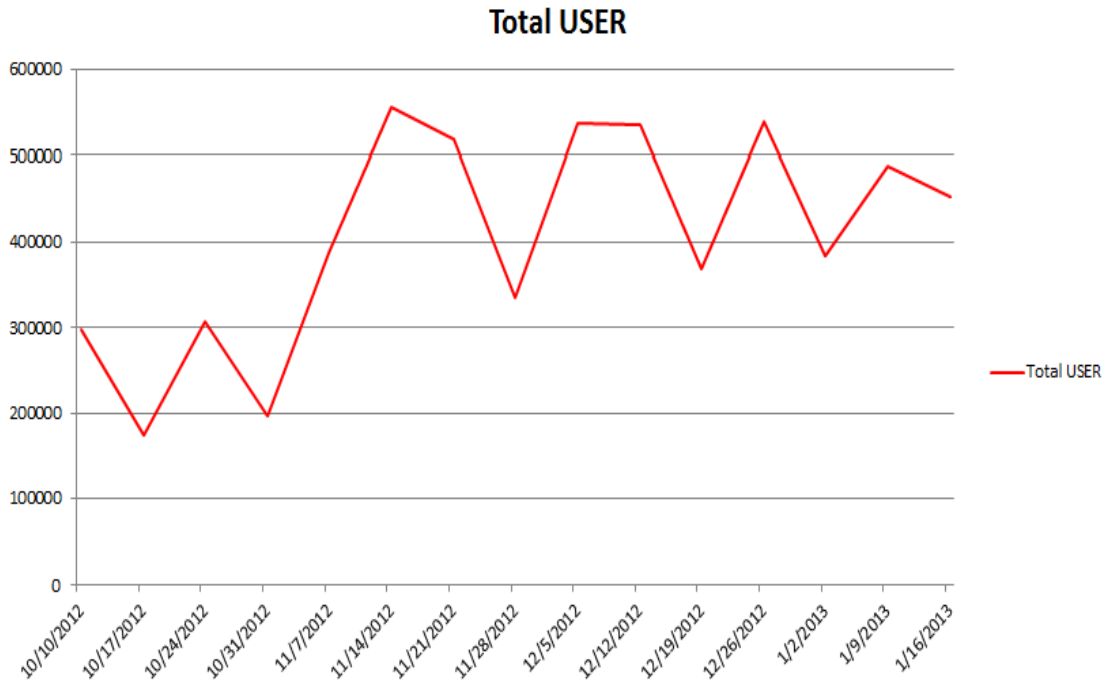

Figure 3.3: Total number of weekly tweet users

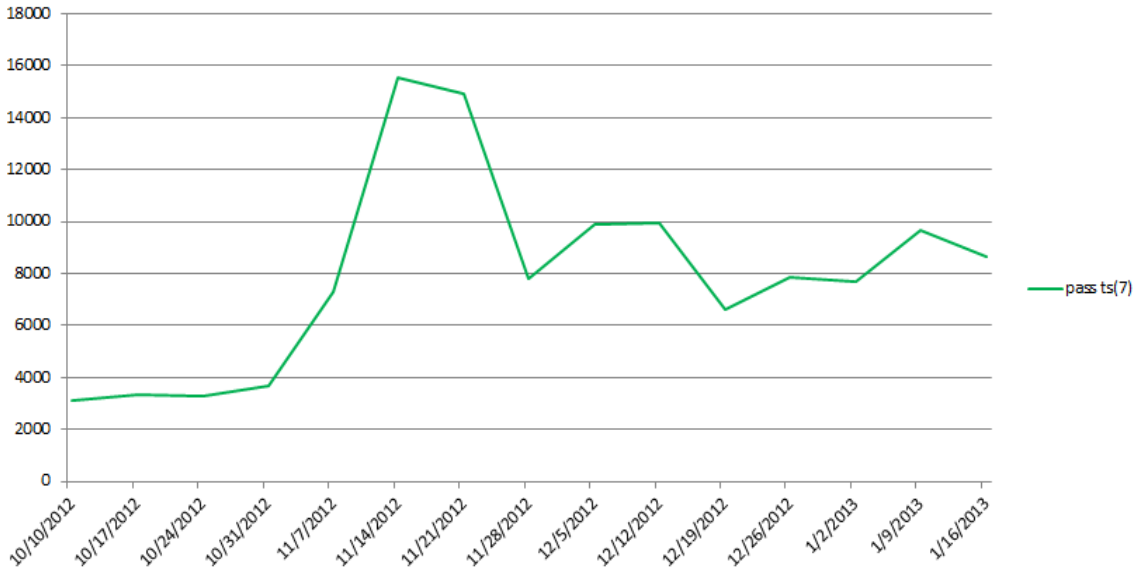

Figure 3.4: Total number of users with 7 or more tweets

3.2 RADICAL AND COUNTER RADICAL USERS

The below charts have the number of radical and counter radical users list. Counter radical users have clearly outnumbered the number of radical users.

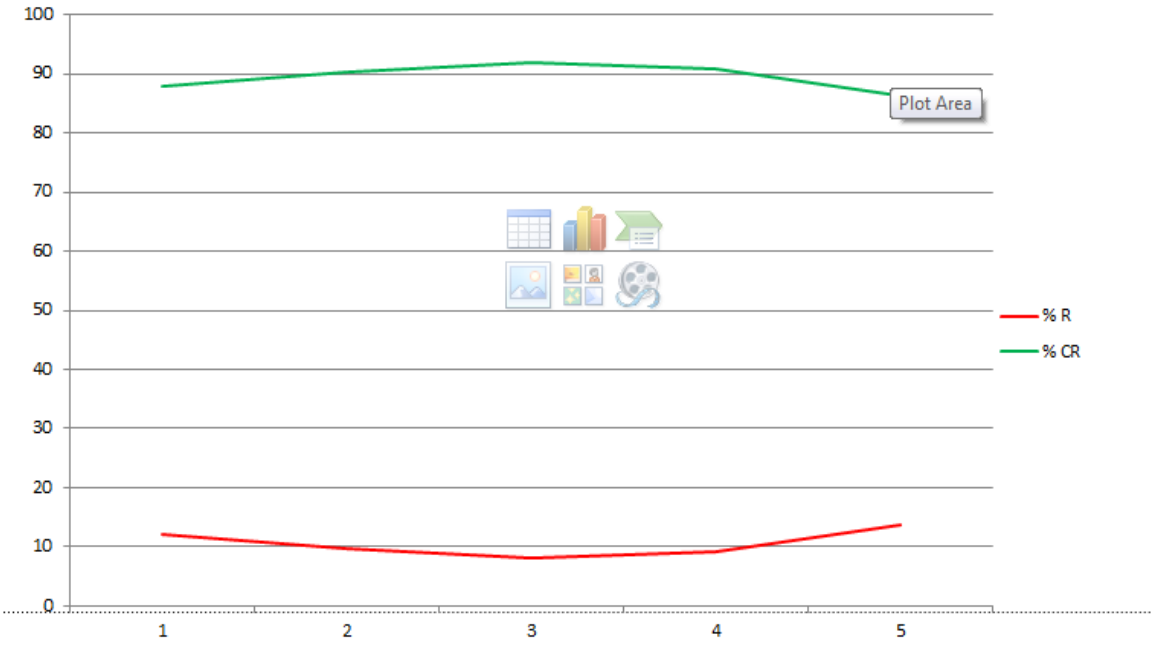

Figure 3.5: Counter radical users vs. Radical users (in %)

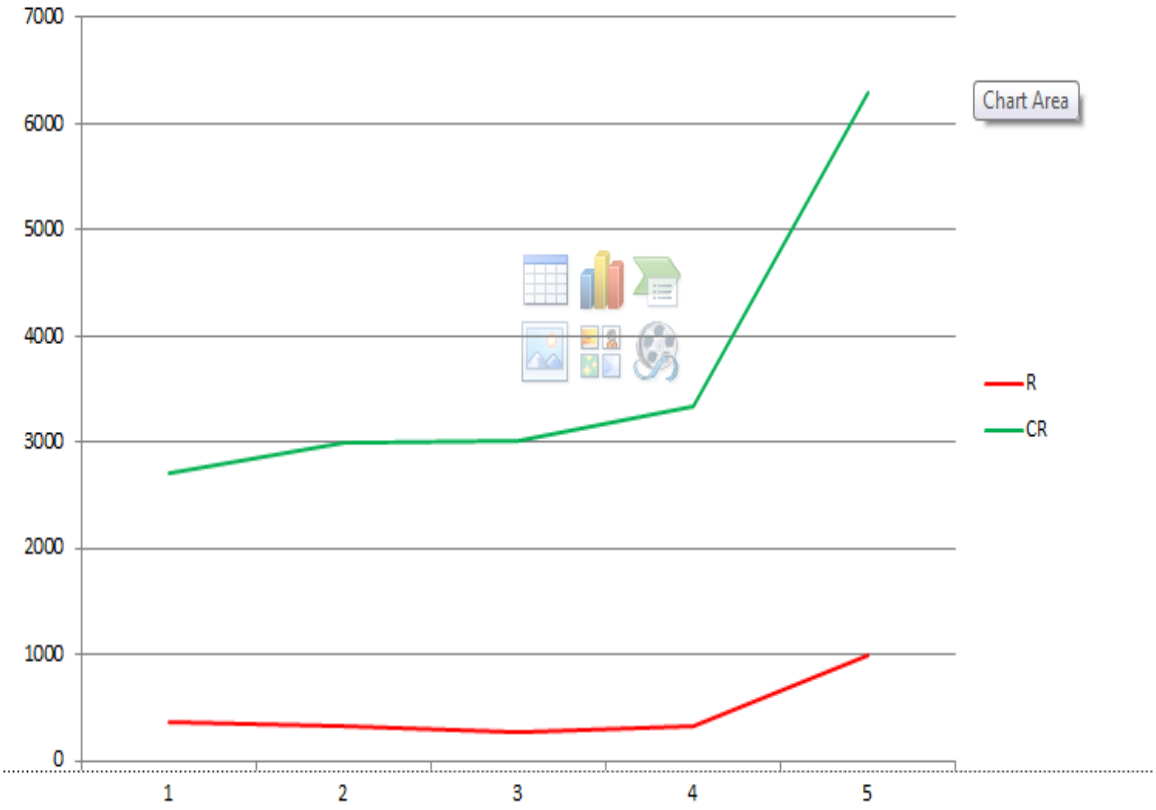

Figure 3.6: Counter radical users vs. Radical users (in numbers)

3.3 ORGANIZATION DISTRIBUTION

The below graph chart explains the users affiliation to a particular organization. In addition to all the organizations, we have two new categories “Unaffiliated_CR” and “Unaffiliated_R”. Unaffiliated_CR users are not affiliated to any organizations and on a counter radical scale. On a similar note Unaffiliated_R users are not affiliated to any organizations and on a radical scale.

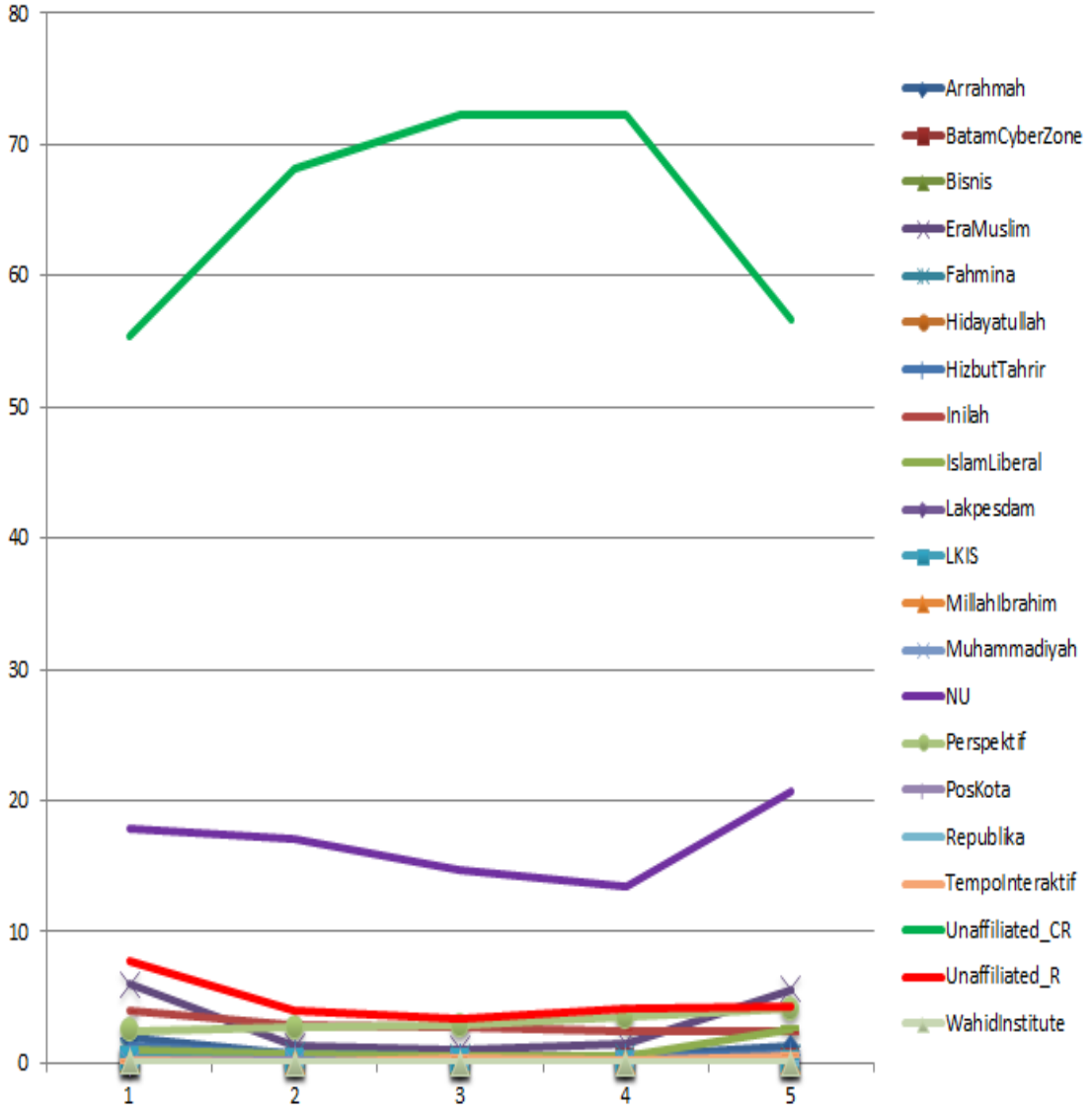

Figure 3.7: Organization distribution list (in %)

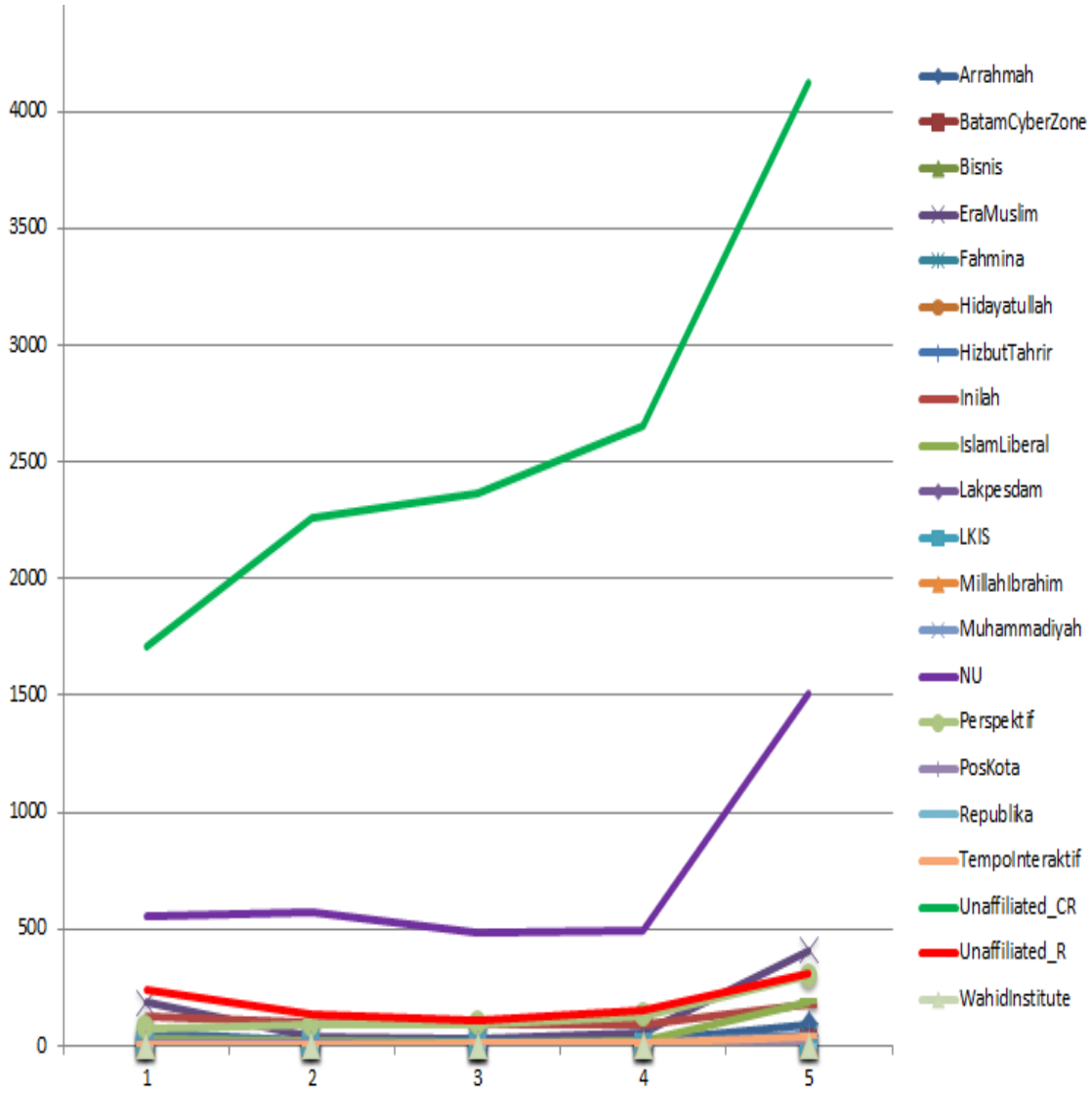

Figure 3.8: Organization distribution list (in numbers)

3.4 SHIFTS IN BEHAVIOR

As we had mapped the users to different organizations on a weekly basis, we were able to predict their shifts and behavior over a period of time. For example, a user who is counter radical in behavior for quite some time could suddenly turn radical because of a sensational event happening in some part of the world. The below graph helps us to understand users shift in behavior over a period of time. For example, radical to counter radical side and vice versa.

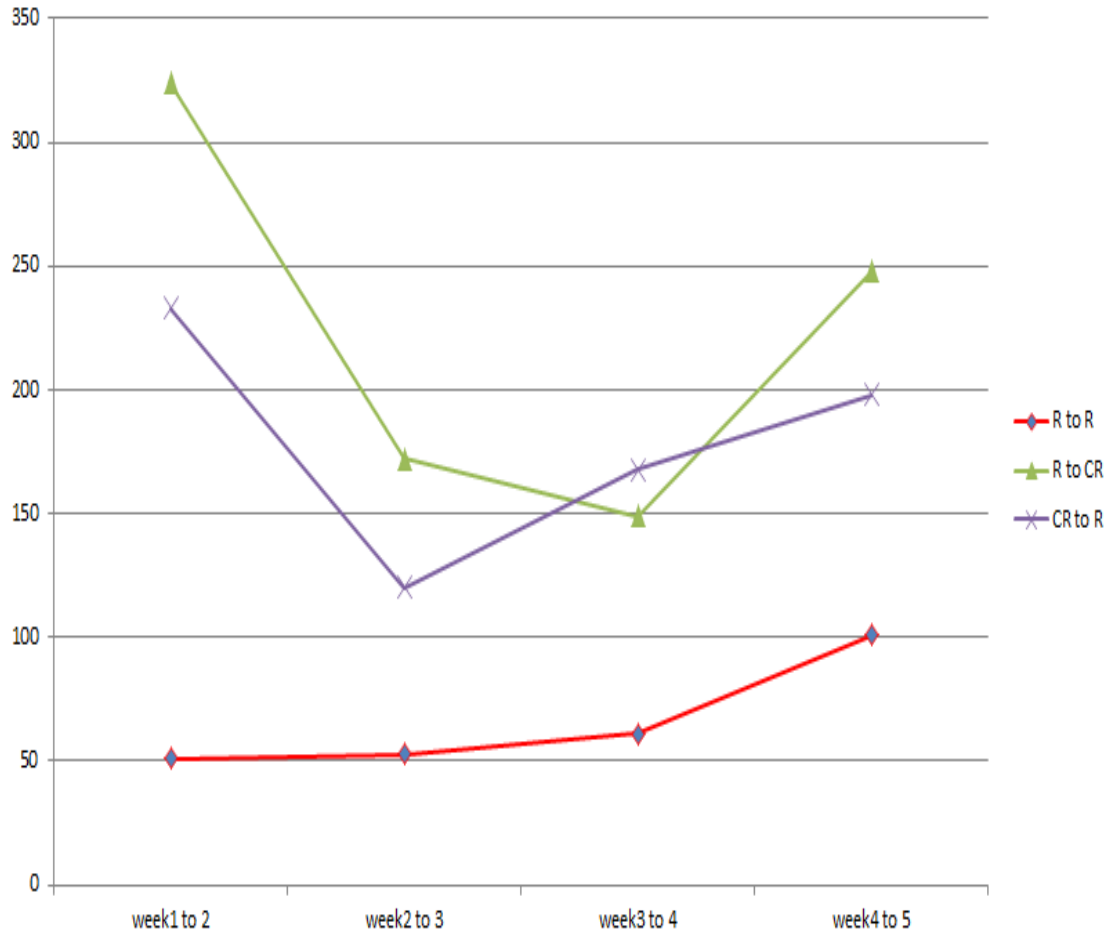

Figure 3.9: Shifts in behavior

3.5 RADICALIZED/ COUNTER RADICALIZED SHIFTS

The user shift can either be radicalized or counter-radicalized. For example, if a user tweet contents were classified to a counter-radical organization in the previous week and the same user contents were classified to a radical organization in the next week, then the user shift is considered as “Radicalized” [10]. On a similar note, if a user tweet contents were classified to a radical organization in the previous week and the same user contents were classified to a counter radical organization in the current week, then the user shift is considered as “Radicalized” [10]. The nature of opinion shifts and the polarities of organizations are shown in Figure 3.10 and “Radicalized/Counter-Radicalized” numbers are shown in Figure 3.11.

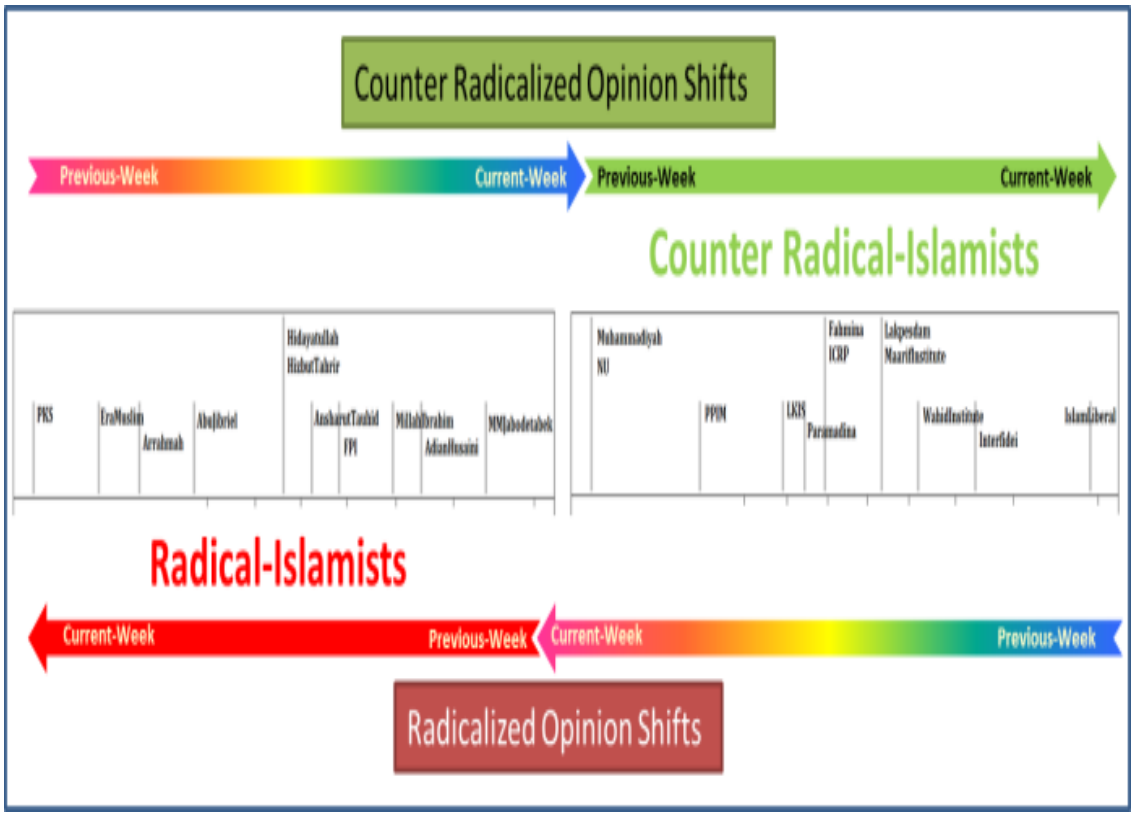

Figure 3.10 Nature of Opinion Shifts and Polarities of Organizations (Image copied from [10])

Figure 3.11: Radicalized/ Counter Radicalized (in numbers)

Chapter 4

SCENARIOS

In this chapter, we will discuss about various scenarios that we were able to track using an interactive web mining dashboard developed at CIPS Research lab, ASU. Some of the scenarios helped us to track radical users and their affiliation with their respective organizations.

1) One of the most famous events at North Sulawesi, Indonesia where a student protested against the security forces [10]. This event happened during October 10, 2012 and October 17 2012 (Figure 4.1).

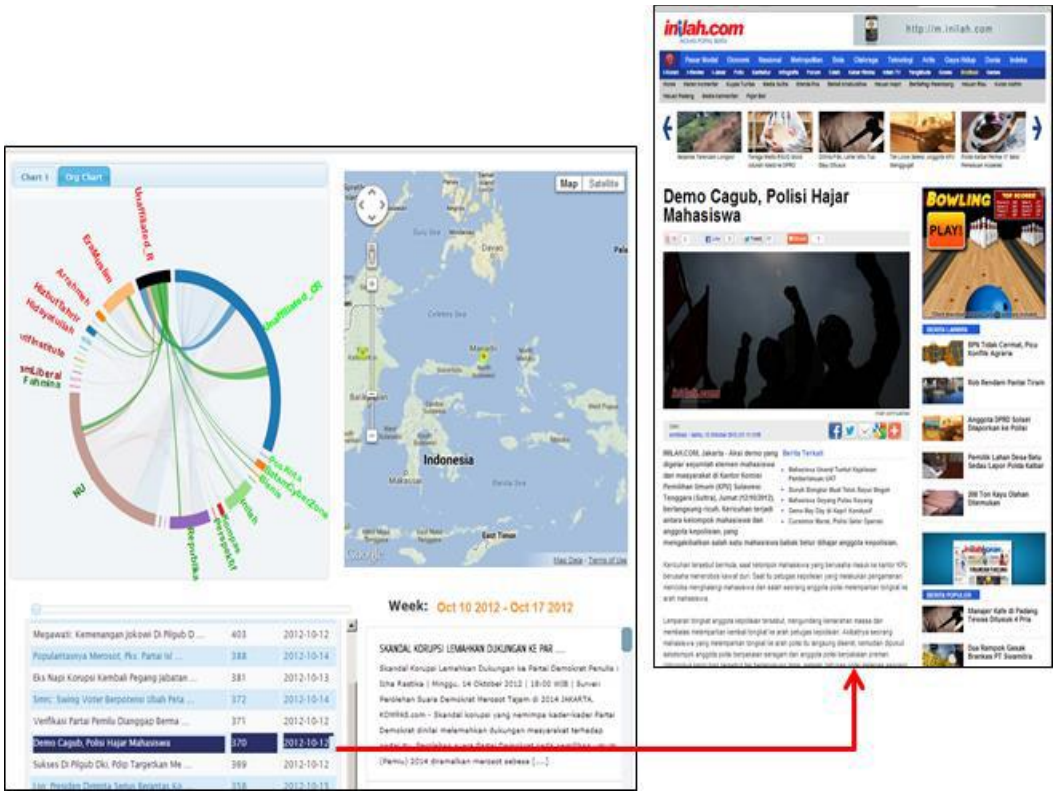

Figure 4.1 Student protests (Note: Image copied from [10])

2) In another interesting scenario, a radical group users exchanging an article that contains information about a missile attack into Israeli territory by a terrorist organization [10]. Figure 4.2 reveals the actual user who exchanged the information in twitter.

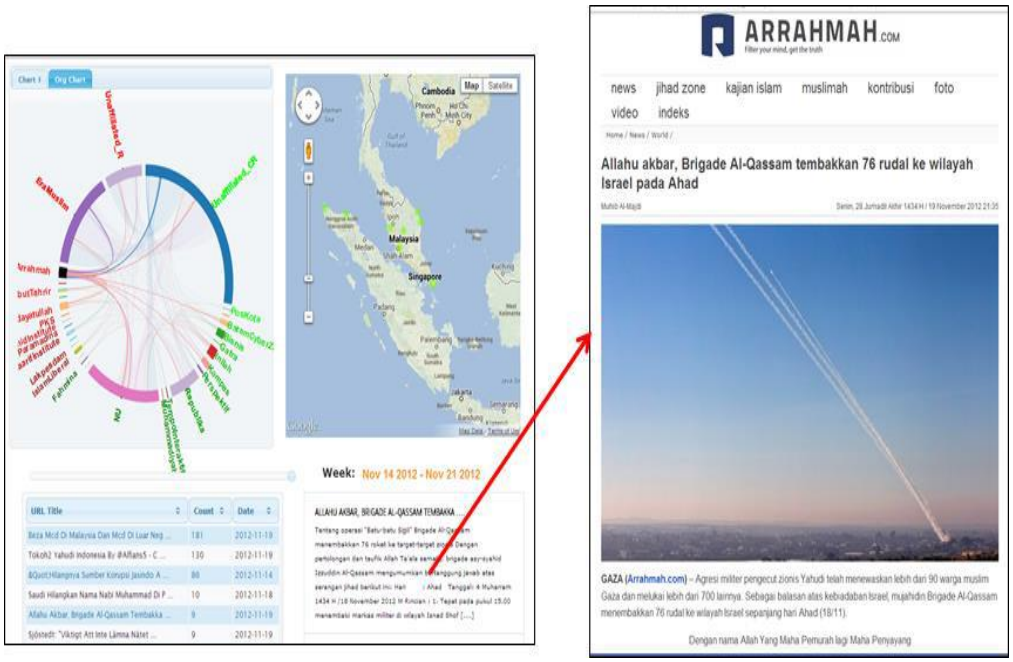

Figure 4.2 Information exchange by radical users (Note: Image copied from [10])

Chapter 5

SUMMARY

In the past decade, social media has taken the world by storm and it allows people to express themselves on virtual world and sometimes enforce their ideologies on others. According to an article [9], Facebook is generating 500 Terabytes of user's data on an average day. Exploiting the textual information collected from Twitter and other news websites, we developed an end-end multi-threaded framework to map users to organizations and thereby discovering hot topics, trends and perspectives.

Initially, we crawled millions of tweets, new articles, events, RSS feeds etc. over a period a time. From the documents crawled, we built linear and logistic regression model using SLEP, an open source sparse learning package. In the second part, we collected, cleaned, tokenized and aggregated all the tweets with respect to every individual user on a weekly basis. Finally, with the help of weekly contents, we computed weekly score and org score for every Individual user. The generated score helped us to analyze the shifts and behavior of the users and discover hot news that causes the shift.

Chapter 6

FUTURE STUDY

Our research study has lot of potentials to dig deep inside and discover new trends. We are currently working on developing a framework for United Kingdom with new enhancements. Some of the research area that we are currently working on is given below.

1) To integrate other social media websites like Facebook, YouTube etc. to our existing system. Since our framework is scalable, mapping news users to one of the existing organization can be achieved.
2) To identify topic and organization specific influential users.
3) To identify sub groups within an organization.
4) To make use of millions of images and videos extracted along with the tweets.
5) To eliminate potential “bot users” aka spammers.
6) To create a location filter using K-Shingles method.

REFERENCES

1) Sukru Tikves, Sedat Gokalp, Mhamed Temkit, Sujogya Banerjee, Jieping Ye, Hasan Davulcu (2012). *Perspective Analysis for Online Debates*. Retrieved from http://www.public.asu.edu/~hdavulcu/FOSINT2012.pdf

2) Sukru Tikves, Sujogya Banerjee, Hamy Temkit, Sedat Gokalp, Hasan Davulcu, Arunaba Sen, Steven Corman, Mark Woodward, Shreejay Nair, Inayah Rohmaniyah, Ali Amin (2012). *A system for ranking organizations using social scale analysis*. Retrieved from http://www.public.asu.edu/~hdavulcu/SNAM12.pdf

3) Kohlschütter, C. (n.d.). Boilerplate Removal and Fulltext Extraction from HTML pages. Retrieved from http://code.google.com/p/boilerpipe/

4) Christian Kohlschütter, Peter Fankhauser, Wolfgang Nejdl (n.d.). Boilerplate Detection using Shallow Text Features. WSDM, Retrieved from http://www.l3s.de/~kohlschuetter/publications/wsdm187-kohlschuetter.pdf

5) Bas van den Beld (n.d.). *How to recognize Twitter bots: 7 signals to look out for.* Retrieved from http://www.stateofsearch.com/how-to-recognize-twitter-bots-6-signals-to-look-out-for/

6) *Longest common substring problem*. Retrieved June 21, 2013, from http://en.wikipedia.org/wiki/Longest_common_substring_problem

7) *Keyword Extraction.* Retrieved from http://www.alchemyapi.com/api/keyword/

8) *NBoilerpipe.* Retrieved from https://github.com/oganix/NBoilerpipe

9) Kern, E. (n.d.). Facebook is collecting your data — 500 terabytes a day. GIGAOM, Retrieved from http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day

10) Nyunsu Kim, Sedat Gokalp, Hasan Davulcu, Mark Woodward (2013). *LookingGlass: A Visual Intelligence Platform for Tracking Online Social Movements* (publication under review)

11) Jun Liu, Shuiwang Ji, and Jieping Ye (2011).*SLEP: Sparse Learning with Efficient Projections*. Retrieved from http://www.public.asu.edu/~jye02/Software/SLEP/manual.pdf

12) Anisha Mazumder, Arun Das, Sedat Gokalp, Nyunsu Kim, Arunabha Sen and Hasan Davulcu (2013). *Spatio-Temporal Signal Recovery from Political Tweets in Indonesia.*(publication under review)